Statistics

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data

Types of Statistics

Descriptive Statistics

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into different categories:

- Measure of central tendency
- Measure of dispersion

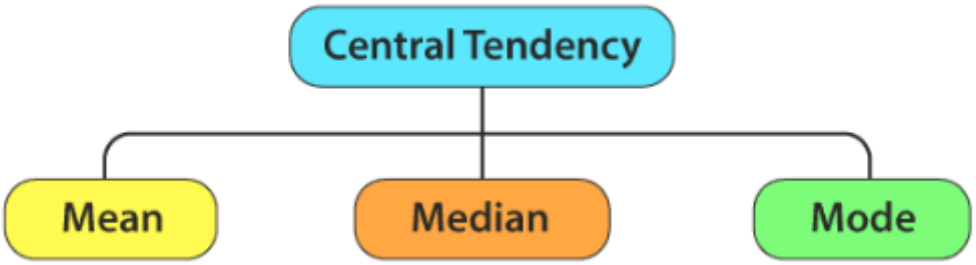
Inferential Statistics

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Population vs Sample

Measures of Central Tendency

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset. It aims to provide an accurate description of the entire data in the distribution.



Mean

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values.

$$\overline{x} = \frac{\sum x}{n}$$

Mean for sample

$$\mu = \frac{\sum x}{n}$$

Mean for population

Salary	15	18	16	14	15	20	24

$$\mu = \frac{15 + 18 + 16 + 14 + 15 + 20 + 24}{7} = \frac{122}{7} = 17.42$$

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Incase of even numbers:

65	55	89	56	35	14	56	55	87	45

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

$$Median\ (even) = \frac{55 + 56}{2} = 55.5$$

Mode

The mode is the most frequent score in our data set

4	3	2	1	1	4	4	5	2	3

$$Mode = 4$$

Measures of Dispersion

The measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

Variance

The average squared deviation from the mean of the given data set is known as the variance. This measure of dispersion checks the spread of the data about the mean.

$$\sigma^2 = \Sigma\frac{(x_i - \mu)^2}{N}$$ variance for population

$$S^2 = \Sigma\frac{(x_i - \bar{x})^2}{n-1}$$ variance for sample

x_i	μ	$(x_i - \mu)$	$(x_i - \mu)^2$
2	3	-1	1
2	3	-1	1
4	3	1	1
4	3	1	1

$$\sigma^2 = \Sigma\frac{(x_i - \mu)^2}{N} = \frac{4}{4} = 1$$

$$S^2 = \Sigma\frac{(x_i - \bar{x})^2}{n-1} = \frac{4}{3} = 1.33$$

Standard Deviation

The square root of the variance gives the standard deviation. Thus, the standard deviation also measures the variation of the data about the mean.

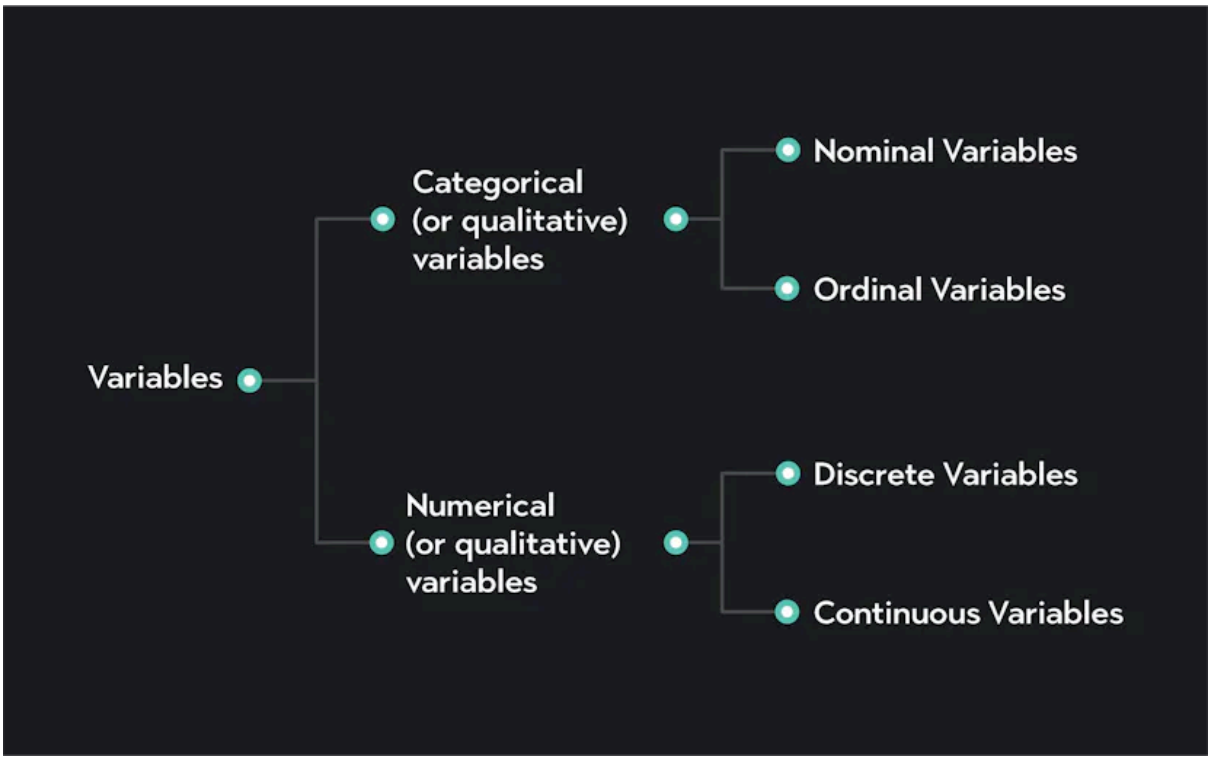
$$S.D = \sqrt{\sigma^2} \quad \text{standard deviation for population}$$

$$S.D = \sqrt{S^2} \quad \text{standard deviation for sample}$$

Variable

A variable is any entity that can take different values.
For example:

- Age is 25
- Gender is male
- Height is 176cm



Quantitative Variable

Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

Discrete Variable

Discrete variables represent counts (e.g. the number of objects in a collection).

For example

- Age is 25
- No of students in a class is 50

Continuous Variable

Continuous variables represent measurable amounts (e.g. water volume or weight).

For example

- Weight is 72.4 kg

Qualitative / Categorical Variable

Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. Qualitative variables are sometimes referred to as categorical variables.

Nominal Variable

Nominal variables are qualitative variables with no inherent ranking based on magnitude or size. Names, nationality, and religion are all nominal variables since there is no way to rank the groups within these categories.

For Example:

- Gender: Male or Female
- Sentiments: Positive, Negative or Neutral

Ordinal Variable

Ordinal variables are qualitative variables where each subgroup can be ranked in order of magnitude.

For Example:

- Education levels: Bachelors, Masters and PhD

Random Variables

A real-valued function, defined over the sample space of a random experiment, is called a random variable. That is, the values of the random variable correspond to the outcomes of the random experiment

Discrete Random Variable

A discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, … and so on

For Example

- Tossing a coin
- Rolling a dice

Continuous Random Variable

A numerically valued variable is said to be continuous if, in any unit of measurement, whenever it can take on the values a and b. If the random variable X can assume an infinite and uncountable set of values, it is said to be a continuous random variable.

For Example

- Today, how many inches rained
- Height of the people

Percentiles and Quartiles

Percentiles

Percentile formula helps in determining the performance of a person in comparison to others. To recall, the percentile is used in tests and scores of a candidate to show where he/she stands with reference to other candidates

$$Percentile\ of\ value\ 'x' = \frac{number\ of\ values\ below\ 'x' * 100}{total\ number\ of\ values}$$

For Example

The scores for student are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91. What is the percentile for score 71?

$$Percentile\ of\ value\ 71 = \frac{6 * 100}{10} = 60$$

Quartiles

Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q1, Q2 and Q3, respectively. Q2 is nothing but the median,

$$Q_1 = [(n + 1)/4]th\ item$$
$$Q_2 = [(n + 1)/2]th\ item$$
$$Q_3 = [3(n + 1)/4]th\ item$$

For Example

Find the quartiles of the following data: 4, 6, 7, 8, 10, 23, 34.

$$Q_1 = [(7 + 1)/4]th\ item = 2nd\ item = 6$$
$$Q_2 = [(7 + 1)/2]th\ item = 4th\ item = 8$$
$$Q_3 = [3(7 + 1)/4]th\ item = 6th\ item = 23$$

Five Number Summary

Lets consider the following data:

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

First, we need to remove outliers

- $Lower\ Fence: Q_1 - 1.5(IQR)$
- $High\ Fence: Q_3 + 1.5(IQR)$
- $IQR = Q_3 - Q_1$
- $Q_1 = \frac{Percentile}{100} * (n + 1) = \frac{25}{100} * 20 = 5th\ item = 3$
- $Q_3 = \frac{Percentile}{100} * (n + 1) = \frac{75}{100} * 20 = 15th\ item = 7$
- $IQR = Q_3 - Q_1 = 7 - 3 = 4$
- $Lower\ Fence: Q_1 - 1.5(IQR) = 3 - 1.5(4) = -3$
- $High\ Fence: Q_3 + 1.5(IQR) = 7 + 1.5(4) = 13$

The new and updated data is

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

- $Minimum$: 1
- Q_1: $[(18 + 1)/4]th\ item = 5th\ item = 3$
- Q_2: $[(18 + 1)/2]th\ item = average\ of\ (9th\ and\ 10th) = 5$
- Q_3: $[3(18 + 1)/4]th\ item = 15th\ item = 7$
- $Maximum$: 9

Covariance and Correlation

Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable

Covariance

Covariance is a measure of how much two random variables changes together. If the variables tend to increase or decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative

Study Hours (X)	**Test Scores (Y)**	$X_I - \overline{X}$	$Y_I - \overline{Y}$	$(X_I - \overline{X})(Y_I - \overline{Y})$
2	**30**	$2 - 6 = -4$	$30 - 64 = -34$	136
4	**50**	$4 - 6 = -2$	$50 - 64 = -14$	28
6	**65**	$6 - 6 = 0$	$65 - 64 = 1$	0
8	**80**	$8 - 6 = 2$	$80 - 64 = 16$	32
10	**95**	$10 - 6 = 4$	$95 - 64 = 31$	124

$$Conv\,(X, Y) = \Sigma\frac{(X_I - \overline{X})(Y_I - \overline{Y})}{n-1} = \frac{136 + 28 + 0 + 32 + 124}{5 - 1} = \frac{320}{4} = 80$$

Advantages

- Quantify the relationship between X and Y

Disadvantage

- Covariance doesnot have a specific limit value

Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear relationship,
- 0 indicates no linear relationship,
- -1 indicates a perfect negative linear relationship.

Pearson Correlation Coefficient

Pearson's correlation measures the linear relationship between two variables. It assumes that both variables are normally distributed and the relationship between them is linear.

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \quad \Sigma(Y_i - \overline{Y})^2}}$$

Study Hours (X)	**Test Scores (Y)**	$X_I - \overline{X}$	$Y_I - \overline{Y}$	$(X_I - \overline{X})(Y_I - \overline{Y})$
2	**30**	$2 - 6 = -4$	$30 - 64 = -34$	136
4	**50**	$4 - 6 = -2$	$50 - 64 = -14$	28
6	**65**	$6 - 6 = 0$	$65 - 64 = 1$	0
8	**80**	$8 - 6 = 2$	$80 - 64 = 16$	32
10	**95**	$10 - 6 = 4$	$95 - 64 = 31$	124

$$r = \frac{136 + 28 + 0 + 32 + 124}{\sqrt{(16 + 4 + 0 + 4 + 16)(1156 + 196 + 1 + 256 + 961)}}$$

$$r = \frac{320}{\sqrt{40 * 4760}} = \frac{320}{\sqrt{190400}} = \frac{320}{436.4} = 0.733$$

Spearman Rank Correlation

Spearman’s correlation measures the monotonic relationship between two variables. It assesses how well the relationship between two variables can be described using a monotonic function (it doesn't have to be linear). Unlike Pearson’s, it does not assume that the variables are normally distributed. Spearman uses the rank of the data rather than the raw values.

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

Here: d is the difference between the ranks of corresponding variables,

Study Hours (X)	Test Scores (Y)	Rank (X)	Rank (Y)
2	**30**	1	1
4	**50**	2	2
6	**65**	3	3
8	**80**	4	4
10	**95**	5	5

Since the ranks are perfectly aligned, di=0d_i = 0di=0 for all data points.

$$\rho = 1 - \frac{6(0)}{5(5^2 - 1)} = 1 - \frac{0}{5(25 - 1)}$$

$$\rho = 1 - 0 = 1$$